

iBus Missing Data Mitigation for Quality of Service Indicators (QSI's)

v2.0 09.12.2013

The purpose of this document is to assist Bus Operators in understanding how missing data is compensated for in iBus QSI results.

This document replaces the 'iBus Data Imputation Process' Issue 1 published in May 2012.

The document is part of a set describing the QSI processes under iBus which also includes 'iBus Data Cleansing for QSIs', 'iBus Data Aggregation for QSIs' and 'iBus QSI Statistics Explained'.

Other relevant reference documents are 'iBus Mileage Cause Codes' and 'iBus Service Control & MTV Coding Guide'.

Contents

1. Introduction	3
2. MTV Missing Trip Coding.....	4
3. High Frequency Data Imputation	5
Overview	5
Stage 1 – Evaluation	5
Stage 2a – Imputation	6
Stage 2b – Exclusion.....	9
Stage 3 – Recalculation of QSI Statistics	9
4. Low Frequency Data Removal.....	10
Overview	10
Stage 1 – Identification of Imputation Cause Codes.....	10
Stage 2a – Time Based Linking.....	10
Stage 2b – Trip Based Linking	11
Stage 3 – Recalculation of QSI statistics.....	13

1. Introduction

- 1.1. There are a number of technical reasons why some actual departure times at QSI points may not be recorded in iBus as observed data. Left unchanged this 'missing data' could have a negative impact on QSI results. The data imputation/removal process is designed to mitigate for missed buses.
- 1.2. The data imputation/removal process is driven by the application of certain detailed cause codes within the MTV Missing Trip Record coding process. See Section 2 for details.
- 1.3. The techniques applied are different for high and low frequency routes.
 - 1.3.1. High frequency routes use data imputation. When a missing trip record is coded via the MTV with an imputable cause code, iBus will aim to impute an observed time for any QSI points falling between the start and end locations of the missing trip record. To do this accurately iBus needs a certain amount of observed data from other buses within a specified time period (typically 3 hours). If sufficient data is not available for a particular route/QSI point/direction, all data within the time period for that route/QSI point/direction will be excluded from the QSI calculations. This is an automatic process. See Section 3 for details.
 - 1.3.2. Low frequency routes use data removal. When a missing trip record is coded via the MTV with an imputable cause code, iBus will ignore such trips for QSI points within the missing section in the 'linking' processes that produce the punctuality statistics and also in the count of expected buses. Unlike high frequency results, no route/QSI point/direction data exclusions arise directly from the low frequency process. See Section 4 for details.
- 1.4. The data imputation/removal process takes place weekly, immediately after the MTV Mileage Coding & Finalisation process.

2. MTV Missing Trip Coding

2.1. Imputation is driven by the application of certain mileage detailed cause codes from either system 'auto-coding' or manual missing trip mileage coding undertaken by Bus Operators.

2.1.1. The imputation process will be triggered by the use of the following mileage detailed cause codes:

Category	Cause	Detailed Cause	Notes
Operated	OP Operated	AC01 Auto coded missing stops at trip start	a b
		AC02 Auto coded missing stops at trip end	a b
		AC03 Auto coded missing stops within trip	a b
		OP04 iBus data not downloaded	a
		OP05 iBus technical errors	a
		OP07 Other	a

a = Unavailable for selection in Service Controller Workstation 'pick list'

b = Unavailable for selection in Missing Trip 'pick list'

2.2. There are three additional operated mileage detailed cause codes that do not lead to imputation:

2.2.1. OP01 'bus on in-service diversion' – the bus did not serve the missing QSI point(s).

2.2.2. OP02 'recovered mileage' - an 'added' trip (not visible in the MTV) will be generated by the 2nd bus for QSI purposes.

2.2.3. OP03 'driver error' - an 'added' trip (not visible in the MTV) will be generated for QSI purposes where an ETM driver login under the wrong trip number is recognised as a scheduled trip. Where ETM driver login fields are not recognised as a valid combination there will be no 'added' trip data.

2.3. Where data is missing for operational reasons this 'lost mileage' will impact on the quality of service provided to passengers and should be reflected in the QSI calculations. Therefore there is no data imputation for missing trip records assigned with lost mileage cause codes.

2.4. For the full list of mileage cause codes refer to the supporting document entitled 'iBus Mileage Cause Codes'.

2.5. For details of when these cause codes should be applied refer to the supporting document entitled 'iBus Service Control & MTV Coding Matrix'.

3. High Frequency Data Imputation

Overview

- 3.1. Imputation is a three stage process.
- 3.2. The first stage involves each dataset (i.e. each route/QSI point/direction and time-period combination) being evaluated against set criteria based on the proportion of imputation cause codes.
- 3.3. Based on the evaluation, the second stage results in either:
 - 3.3.1. The imputation process being run for the dataset, or
 - 3.3.2. All hourly totals within a dataset being marked as excluded.
- 3.4. In the final stage the QSI statistics are recalculated.

Stage 1 – Evaluation

- 3.5. For each data set the proportion of imputation cause codes is calculated by the following equation:
$$\% \text{ imputation cause codes} = \text{Sum (imputation cause codes)} / \text{Sum (observed)} * 100$$

where 'imputation cause codes' are scheduled QSI point events without an observation and with an imputable cause code (see Section 2 for details).
 - 3.5.1. If the percentage of imputation cause codes is greater or equal to a threshold of 20%, the imputation process will not run and all hourly totals for the dataset will be marked as excluded.
 - 3.5.2. If the percentage of imputation cause codes is less than 20% for a dataset, the imputation process will run for that dataset.
 - 3.5.3. If the formula leads to a 'divided by zero' scenario, the imputation process will not run and all hourly totals for the dataset will be marked as excluded.

Example 1 shows a QSI point with 15 observed buses for the time period 07:00 to 10:00. There are 5 expected buses with no observed data present, 2 of which have been assigned an imputable cause code (OP04 & OP05).

- The percentage of imputation cause codes is $2 / 15 = 13\%$.
- As the percentage is less than 20% the imputation process will run.

QSI 1 Scheduled Time	QSI 1 Observed Time	QSI 1 Cause Code	% Imputable Codes	Decision
07:00	07:02			
07:10	07:09			
07:20	07:22			
07:30	#	OP04 (imputable cause code)		Impute time
07:40	07:35			
07:50	07:48			
08:00	08:03			
08:10	08:10			
08:20	08:18			
08:30	08:26			
08:40	#	OP05 (imputable cause code)		Impute time
08:50	08:47			
09:00	09:01			
09:10	09:09			
09:20	09:20			
09:30	#	OD02 (non-imputable cause code)		
09:40	#	ON05 (non-imputable cause code)		
09:45	#	OP02 (non-imputable cause code)		
09:50	09:46			
09:55	09:52			
Sum (observed) = 15		Sum (imputable cause codes = 2	2 / 15 = 13%	No exclusion

Example 2 shows a QSI point with 15 observed buses for the time period 07:00 to 10:00. There are 5 expected buses with no observed data present, all of which have been assigned an imputable cause code (OP04 & OP05).

- The percentage of imputation cause codes is $5 / 15 = 33\%$.
- As the percentage is greater than 20% the imputation process will not run and the hourly totals for the dataset are marked as excluded.

QSI 1 Scheduled Time	QSI 1 Observed Time	QSI 1 Cause Code	% Imputable Codes	Decision
07:00	07:02			Exclude
07:10	07:09			Exclude
07:20	07:22			Exclude
07:30	#	OP04 (imputable cause code)		Exclude
07:40	07:35			Exclude
07:50	07:48			Exclude
08:00	08:03			Exclude
08:10	08:10			Exclude
08:20	08:18			Exclude
08:30	08:26			Exclude
08:40	#	OP05 (imputable cause code)		Exclude
08:50	08:47			Exclude
09:00	09:01			Exclude
09:10	09:09			Exclude
09:20	09:20			Exclude
09:30	#	OP04 (imputable cause code)		Exclude
09:40	#	OP05 (imputable cause code)		Exclude
09:45	#	OP05 (imputable cause code)		Exclude
09:50	09:46			Exclude
09:55	09:52			Exclude
Sum (observed) = 15		Sum (imputable cause codes = 5	5 / 15 = 33%	Exclude data set

Stage 2a – Imputation

- 3.6. The imputation process is applied at time period level. Each route/direction/QSI point combination is treated separately for each time period. A mathematical algorithm is applied to the observed headways for other buses in the time period to derive an imputed 'observation' time for a missing bus. An imputed 'observation' time not intended to replicate when the 'missing' bus actually departed but is a statistical representation used to split a single headway into two, thus providing a more accurate AWT result for that route, direction, QSI point and time period.
- 3.7. There may be occasions where, due to the nature of the headways, imputed 'observation' times are earlier than observed times for the previous stop (or later than the subsequent stop). This is because the imputation algorithm uses inputs from each QSI point in isolation.
- 3.8. The following steps are used:
 - 3.8.1. All observed headways within the time period are allocated a 'class' where each class is an interval of 60 seconds. So, for example, an observed headway of 135 seconds will be placed in the class of 120 to 180 seconds. Intervals are 0 to 60 seconds, 60 to 120 seconds, 120 to 180 seconds, 180 to 240 seconds, etc. For each class the frequencies are derived.
 - 3.8.2. A discrete probability distribution is created, by multiplying the frequency of occurrence in each interval by the total number of observations. So for example if there were 3 headways within the 120 to 180 second interval, and 10 observations in total, then the probability in that interval would be 0.3.
 - 3.8.3. The cumulative headway probability is then calculated and values up to and including 0.9 (the 90th percentile) are multiplied by the intervals mid range value (in seconds). The sum of these values is the 'imputed headway'.
- 3.9. The algorithm is designed to reduce any 'error' between actual and estimated observed times as much as possible. The 'imputed headway value' is slightly less than the average of observed headways within the time period (at the 90th percentile of the discrete probability distribution). This 'imputed headway value' is then added to the observed time of a previous scheduled trip to derive the imputed 'observation' time for the missing trip.
 - 3.9.1. Where two observed times exist for a scheduled trip (i.e. one 'added' trip and one 'validated' trip) the observed time of the validated trip will always be used.
 - 3.9.2. Where the time to be imputed relates to the first trip of the day (for which there is no previous observation time) the time imputed will be the scheduled departure time at that location.

Example 3 below provides a demonstration of how imputation is applied.

- Observed headways are assigned to time periods based on when they end. So in the example, headways for the 07:00-10:00 time-period are shown in blue, and include the headway from 06:51 to 07:01, but not the headway from 09:49 to 10:01.
- The imputed time for trip 47 is 11:01:35, being the observed time of trip 45 (10:51) + the 90th percentile headway value (10m 35s). The observed time for trip 45 is used even though this is out of sequence with trips 41 and 43.
- The 90th percentile headway value for trip 61 is taken from the 10:00-13:00 calculation as the scheduled departure time is 12:48 even though the preceding and following observed times are both within the 13:00-16:00 time-period.
- Where there are no observed departure times for consecutive trips in the same direction, the 90th percentile headway value for the time period is multiplied by the number of consecutive trips since the last observed time. Where there is an imputation cause code following two consecutive non-imputation cause codes the same rule applies.
- In the example, the imputed time for trip 7 is 07:18:33, being the observed time of trip 5 (07:05) + the 90th percentile headway value (13m 33s).
- The imputed time for trip 9 is 07:32:06, being the observed time of trip 5 (07:05) + twice the 90th percentile headway value (2* 13m 33s).
- The imputed time for trip 25 is 09:14:39, being the observed time of trip 19 (08:34) + three times the 90th percentile headway value (3* 13m 33s).

Trip	Scheduled time	Observed time	Cause Code	Sorted observed times	Observed headways (mins)
1	06:48	06:51		06:51	
3	07:00	07:01		07:01	00:10
5	07:12	07:05		07:05	00:04
7	07:24	07:18:33	OP05	07:47	00:42
9	07:36	07:32:06	OP05	07:49	00:02
11	07:48	07:47		08:14	00:25
13	08:00	07:49		08:26	00:12
15	08:12	08:14		08:34	00:08
17	08:24	08:26		09:21	00:47
19	08:36	08:34		09:34	00:13
21	08:48		TR01	09:49	00:15
23	09:00		TR01	10:01	00:12
25	09:12	09:14:39	OP05	10:09	00:08
27	09:24	09:21		10:26	00:17
29	09:36	09:34		10:36	00:10
31	09:48	09:49		10:51	00:15
33	10:00	10:01		11:02	00:11
35	10:12	10:09		11:11	00:09
37	10:24	10:26		11:18	00:07
39	10:36	10:36		11:53	00:35
41	10:48	11:02		11:56	00:03
43	11:00	11:11		12:10	00:14
45	11:12	10:51		12:36	00:26
47	11:24	11:01:35	OP05	13:03	00:27
49	11:36	11:18		13:09	00:06
51	11:48	11:53		13:18	00:09
53	12:00	11:56			
55	12:12	12:10			
57	12:24	12:36			
59	12:36	13:03			
61	12:48	13:13:35	OP05		
63	13:00	13:09			
65	13:12	13:18			

Calculation of 90 th percentile headway value, 07:00 – 10:00					
Headway (mins)	Mid range (secs)	prob	cumulative	Bin mid-range*prob	
2	150	0.1	0.1	15	
4	270	0.1	0.2	27	
8	510	0.1	0.3	51	
10	630	0.1	0.4	63	
12	750	0.1	0.5	75	
13	810	0.1	0.6	81	
15	930	0.1	0.7	93	
25	1530	0.1	0.8	153	
42	2550	0.1	0.9	255	
47	2850	0.1	1.0	813	
Value =				13m 33s	

Calculation of 90 th percentile headway value, 10:00 – 13:00					
Headway (mins)	Mid range (secs)	prob	cumulative	Bin mid-range*prob	
3	210	0.08	0.08	16	
7	450	0.08	0.15	35	
8	510	0.08	0.23	39	
9	570	0.08	0.31	44	
10	630	0.08	0.38	48	
11	690	0.08	0.46	53	
12	750	0.08	0.54	58	
14	870	0.08	0.62	67	
15	930	0.08	0.69	72	
17	1050	0.08	0.77	81	
26	1590	0.08	0.85	122	
27	1650	0.08	0.92		
35	2130	0.08	1.00	635	
Value =				10m 35s	

Stage 2b – Exclusion

- 3.10. Where the proportion of imputation cause codes is above the 20% threshold all hourly totals within the dataset will be marked as excluded and are shown as 'imputation exclusions'.

Stage 3 – Recalculation of QSI Statistics

- 3.11. Once data exclusions or new observed departure times have been generated, QSI statistics for hourly totals are re-calculated.

4. Low Frequency Data Removal

Overview

- 4.1. It is not appropriate to impute 'observation' times for low frequency routes. Unlike headways on high frequency routes even a one second variance on the imputed time can mean, for example, the difference between being assessed as 'late' instead of 'on-time'. Consequently a different approach is taken and scheduled times are removed from the linking process itself.
- 4.2. Linking is performed daily during staging. Once Bus Operator mileage claims have been finalised, linking is performed for a final time using the following stages:
 - 4.2.1. Identify whether any imputation cause codes have been assigned.
 - 4.2.2. Where such cause codes are found:
 - For time-based linking, perform linking where scheduled times with imputation cause codes are skipped during the linking process, or
 - For trip-based linking, change 'Non-Arrival' status to "NULL" where scheduled times have imputation cause codes.
 - 4.2.3. Recalculate the QSI statistics, using a new expected buses denominator.

Stage 1 – Identification of Imputation Cause Codes

- 4.3. Each dataset (each route/QSI point/direction and service date combination) is evaluated to identify whether any imputation cause codes have been assigned (see Section 2 for details).

Stage 2a – Time Based Linking

- 4.4. Scheduled buses that are assigned imputation cause codes are ignored during the linking process.
- 4.5. Linking is performed by working through each scheduled bus at each QSI point and attempting to fit an observed bus to it (as 'on-time', 'late', 'early' or 'non-arrival'). Where a scheduled bus has an imputation cause code assigned this process is skipped and the 'link status' field becomes "NULL".
- 4.6. Time linking is demonstrated in the following examples. . .

Example 4 shows how initial linking is undertaken pre imputation (for comparative purposes only). Here, the scheduled bus at 07:40 with an imputation cause code is linked to an observed time at 07:39 as 'on-time'.

- There are 9 linked buses out of the expected 12 (75%).

Trip	QSI 1	QSI 1	Cause Code	Link Status
	Scheduled Time	Observed Time		
1	07:10	07:05:00		Early
3	07:25	07:39:00		Non-arrival
5	07:40	#	OP04 (imputable cause code)	On-time
7	07:55	07:53:09		On-time
9	08:10	08:11:03		On-time
11	08:25	#	OP05 (imputable cause code)	Non-arrival
13	08:40	08:45:03		Late
15	08:55	08:55:15		On-time
17	09:10	09:23:01		Non-arrival
19	09:25	09:43:08		On-time
21	09:40	09:54:45		On-time
23	09:55			On-time

Example 5 – shows how re-linking for the same set of scheduled and observed times is done post imputation with selective removal of scheduled buses from the linking process. The scheduled bus at 07:40 with an imputation cause code is no longer linked to observed time 07:39 (the red arrow indicates the link that was made in example 4). Scheduled bus 07:25 now links with observed time 07:39 to create a 'Late' status and the scheduled bus at 07:40 now has no link status.

- With the schedule buses with imputation cause codes removed from the linking process there are now 9 linked buses out 10 (90%).

Trip	QSI 1	QSI 1	Cause Code	Link Status
	Scheduled Time	Observed Time		
1	07:10	07:05:00		Early
3	07:25	07:39:00		Late
5	07:40 (removed)	#	OP04 (imputable cause code)	
7	07:55	07:53:09		On-time
9	08:10	08:11:03		On-time
11	08:25 (removed)	#	OP05 (imputable cause code)	
13	08:40	08:45:03		Late
15	08:55	08:55:15		On-time
17	09:10	09:23:01		Non-arrival
19	09:25	09:43:08		On-time
21	09:40	09:54:45		On-time
23	09:55			On-time

Stage 2b – Trip Based Linking

- 4.7. Where linking is performed by matching scheduled and observed trip numbers, it is not possible for observed times to be linked to different scheduled times (unlike time-based linking) and a different approach is employed. Where scheduled times have imputation cause codes, the link status is changed from 'Non-Arrival' to 'NULL'
- 4.8. Trip Linking is demonstrated in the following examples.

Example 6 shows how initial linking is undertaken pre imputation (for comparative purposes only).

Scheduled Trip #	Scheduled Time	Observed Trip #	Observed Time	Cause Code	Link Status
1	15:53	1	15:59:00		Late
3	16:09	3	16:00:00		Early
		5	16:21:00		
5	16:24	5	16:26:00		On-time
7	16:39			OP05 (imputable)	Non-arrival
9	16:54	9	16:50:00		Early
11	17:09	11	16:52:00		Early
13	17:24	13		OP04 (imputable)	Non-arrival
15	17:39	15	17:29:00		Early
17	17:54	17		ST01 (non-imputable)	Non-arrival
19	18:09	19	17:56:00		Early
			Total	Sum (On-time)	1
				Sum (Early)	5
				Sum (Late)	1
				Sum (Non-arrival)	3
				Sum (All link status)	10
			Hourly totals	On-Time	10.0%
				Early	50.0%
				Late	10.0%
				Non-arrival	30.0%

Example 7 shows the same set of scheduled and observed times post imputation with the selective removal of ‘non-arrival’ link status and recalculation of hourly totals. This shows how the link status is changed from ‘Non-Arrival’ to ‘NULL’ and the hourly totals are recalculated:

- either with a changed denominator (Expected buses – imputation cause codes),
- or sum(on-time + late + early + non-arrival).

Scheduled Trip #	Scheduled Time	Observed Trip #	Observed Time	Cause Code	Old Link Status	New Link Status
1	15:53	1	15:59:00		Late	Late
3	16:09	3	16:00:00		Early	Early
		5	16:21:00			
5	16:24	5	16:26:00		On-time	On-time
7	16:39			OP05 (imputable)	Non-arrival	
9	16:54	9	16:50:00		Early	Early
11	17:09	11	16:52:00		Early	Early
13	17:24	13		OP04 (imputable)	Non-arrival	
15	17:39	15	17:29:00		Early	Early
17	17:54	17		ST01 (non-imputable)	Non-arrival	Non-arrival
19	18:09	19	17:56:00		Early	Early
				Total	Sum (On-time)	1
					Sum (Early)	5
					Sum (Late)	1
					Sum (Non-arrival)	1
					Sum (All link status)	8
				Hourly totals	On-Time	12.5%
					Early	62.5%
					Late	12.5%
					Non-arrival	12.5%

Stage 3 – Recalculation of QSI statistics

- 4.9. Once linking is complete, the hourly totals statistics are generated. The difference with pre-imputation is that ‘expected buses’ has changed.